3

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-348758
(43)Date of publication of application : 22.12.1994

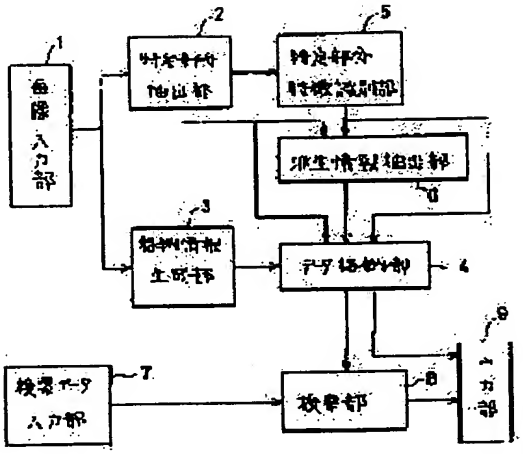
(51)Int.Cl. G06F 15/40
G06F 15/62

(21)Application number : 05-133746
(22)Date of filing : 04.06.1993

(71)Applicant : TOSHIBA CORP
(72)Inventor : KUROSAWA YOSHIAKI
TANAKA HISAKO
MATSUMURA YOSHIKUNI

(54) DEVICE AND METHOD FOR RETRIEVING DOCUMENT INFORMATION

(57)Abstract:
PURPOSE: To provide a device and method for retrieving document information to retrieve a desired document with information expressing the external appearance of the document as a retrieval key.
CONSTITUTION: The specified part (such as a background part, character part, ruled line part of a table, drawing part or photograph part) of inputted image data is extracted (2), the feature (such as a size, position, kind or color) is discriminated, information expressing the external appearance of the document (such as the color of paper, paper quality, spot, color of characters, character type, character kind, writer, writing tool, font, amount of writing in margins, longitudinal and lateral lengths of documents, size/density of characters, kind of paper or positional relation of drawings or photographs) is extracted (6), and this information expressing the external appearance of the document is stored in a data storage part (4) taking correspondence with a document picture or one of recognition result of it. At the time of retrieval, the desired document is retrieved by using this information expressing the external appearance of the document.



LEGAL STATUS

[Date of request for examination]	30.03.2000
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	3142986
[Date of registration]	22.12.2000
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号

特開平6－348758

(43)公開日　平成６年(1994)12月22日

(51)Int.Cl.[5]		識別記号	庁内整理番号	Ｆ Ｉ	技術表示箇所
Ｇ０６Ｆ	15/40	５３０　Ｌ	9194－5L		
	15/62	３３０　Ｇ	8125－5L		

審査請求　未請求　請求項の数５　ＯＬ　（全 14 頁）

(21)出願番号　特願平5－133746

(22)出願日　平成５年(1993)６月４日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　黒沢　由明
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(72)発明者　田中　久子
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(72)発明者　松村　善邦
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

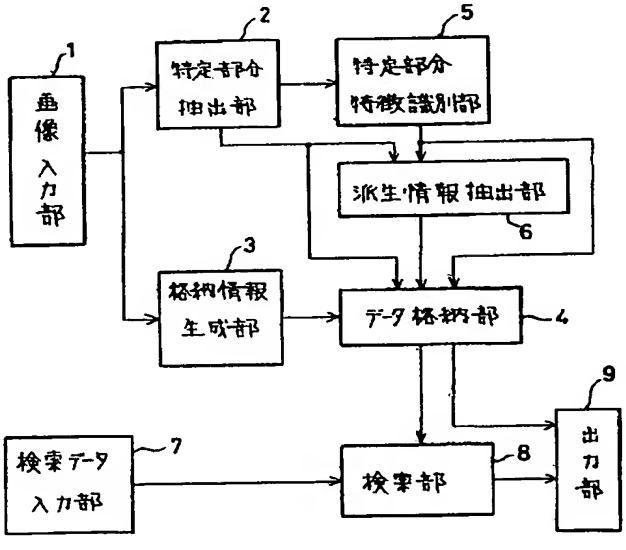
(74)代理人　弁理士　則近　憲佑

(54)【発明の名称】　文書情報検索装置及び方法

(57)【要約】

【目的】　文書の外観を表す情報を検索キーとして所望の文書を検索できる文書情報検索装置及び方法の提供を目的とする。

【構成】　入力された画像データの特定部分（バックグラウンド部、文字部、表の罫線部、図面部、写真部等）を抽出し（２）、その特徴（大きさ、位置、種類、色等）を識別し（５）、文書の外観を表す情報（紙の色、紙質、シミ、文字の色、文字タイプ、文字種、筆記者、筆記具、フォント、余白への書き込みの量、書類の縦横、文字の大きさ・密度、用紙の種類、図面や写真の位置関係等）を抽出し（６）、この文書の外観を表す情報を文書画像あるいはこれを認識処理したものに対応づけてデータ格納部（４）に格納する。検索時には、この文書の外観を表す情報を用いて所望の文書を検索する。



1

【特許請求の範囲】

【請求項１】　入力された文書画像あるいはこの文書画像に処理を施した結果を文書として記憶する記憶手段と、記憶された文書を文書名あるいはキーワードにより指定する指定手段と、指定された文書を出力する出力手段とを備える文書情報検索装置において、

前記入力手段により入力された文書画像から文書の外観的な特徴を表す情報を抽出する手段と、この手段により抽出された情報を前記記憶手段に記憶された文書に対応付けて記憶する手段とを具備し、且つ前記指定手段による文書の指定に替わりあるいは前記指定手段による文書の指定に加えて、所望の文書の外観的な特徴を示す情報を検索キーとして入力し、入力された検索キーと文書に対応付けて記憶された前記情報との照合を行って、出力すべき文書を決定する手段を備えたことを特徴とする文書情報検索装置。

【請求項２】　入力された文書画像あるいはこの文書画像に処理を施した結果を文書として記憶するステップと、

入力された文書画像から抽出される文書の外観的な特徴を表す情報を前記文書に対応付けて記憶するステップと、

所望の文書の外観的な特徴を示す情報を検索キーとして入力するステップと、

入力された検索キーと記憶された前記情報との照合を行うステップと、

この照合結果に従い記憶された前記文書の中から対応する文書を出力するステップとを備えたことを特徴とする文書情報検索方法。

【請求項３】　媒体に定着した文書画像を入力する手段と、

入力された文書画像あるいはこの文書画像に処理を施した結果を文書として記憶する手段と、

入力された文書画像から前記媒体の特徴を表す情報を抽出し、この情報を前記文書に対応付けて記憶する手段と、

所望の文書画像がどのような媒体に定着していたかを示す情報を検索キーとして入力する手段と、

入力された検索キーと記憶された前記情報との照合を行う手段と、

この照合結果に従い記憶された前記文書の中から対応する文書を出力する手段とを備えたことを特徴とする文書情報検索装置。

【請求項４】　媒体に定着した文書画像を入力する手段と、

入力された文書画像あるいはこの文書画像に処理を施した結果を文書として記憶する手段と、

入力された文書画像から前記媒体に定着した物質の特徴を表す情報を抽出し、この情報を前記文書に対応付けて記憶する手段と、

2

所望の文書画像がどのような物質により媒体に定着していたかを示す情報を検索キーとして入力する手段と、

入力された検索キーと記憶された前記情報との照合を行う手段と、

この照合結果に従い記憶された前記文書の中から対応する文書を出力する手段とを備えたことを特徴とする文書情報検索装置。

【請求項５】　媒体に定着した文書画像を入力する手段と、

入力された文書画像あるいはこの文書画像に処理を施した結果を文書として記憶する手段と、

入力された文書画像から前記媒体上に表された情報のイメージとしての特徴を表す情報を抽出し、この情報を前記文書に対応付けて記憶する手段と、

所望の文書画像がどのようなイメージで媒体上に表されていたかを示す情報を検索キーとして入力する手段と、

入力された検索キーと記憶された前記情報との照合を行う手段と、

この照合結果に従い記憶された前記文書の中から対応する文書を出力する手段とを備えたことを特徴とする文書情報検索装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、文書画像あるいはこれを認識処理した結果をファイルする装置において、所望の文書を検索するための文書情報検索装置及び方法に関する。

【０００２】

【従来の技術】従来、文書画像をスキャナで読み込みこの画像情報を蓄積することにより、紙の書類のファイルに代わるものを電子的に実現するファイリング装置が提供されてきた。しかし、膨大な量の文書の蓄積がなされるようになると、ファイルしたは良いが、その文書に付されたキーワードを忘れてしまい、膨大な文書の中から所望のものを探し出すのに多大な労力がかかることとなり、このようなファイリング装置は非常に使いづらいものになっていた。

【０００３】紙の書類であれば、その外観的な特徴から比較的簡易に、例えば「コーヒーのしみが付いていた文書」のような人間の記憶を基に、探し出すことができるが、電子化される際にこのような情報は余分なものとして捨てられてしまうため、従来の電子ファイリング装置では、紙の場合のように人間の自然な思い出し方で所望の文書を探すことができない。

【０００４】特に、文書画像に対して文書構造解析や文字認識等の処理を施したものをファイルする装置では、認識処理の際に、認識率を上げるためしみのようなノイズは除去されてしまう。そして、検索時に表示される文書は、認識処理が施された後のものであって、人間の記憶に残っている元の文書画像とは外観上異なるものにな

ってしまうため、一見して所望の文書であるかどうかを判別できない。

【０００５】

【発明が解決しようとする課題】このように、従来の電子ファイリング装置には、人間の記憶に残り易い、文書の外観的な特徴によっては、ファイルされた文書情報を検索することができないという問題点があった。

【０００６】本発明はこの点に鑑みてなされたもので、文書の外観的な特徴を検索キーとして所望の文書を検索できる文書情報検索装置及び方法を提供することを目的とする。

【０００７】

【課題を解決するための手段】本発明に係る文書情報検索装置及び方法は、入力された文書画像から抽出される文書の外観的な特徴を表す情報を入力された前記文書画像あるいはこれを認識処理したものに対応付けて記憶しておき、文書の外観的な特徴を示す情報が検索キーとして入力されると、入力された検索キーと記憶された前記情報との照合を行い、この照合結果に従い対応する前記文書画像あるいは認識処理されたものを出力することを特徴とする。

【０００８】この文書の外観的な特徴を表す情報を用いた検索を、文書名あるいはキーワードを指定する検索と併用することもできる。文書の外観的な特徴を表す情報は、以下の３つの種類に大別される。（１）文書画像が定着していた媒体の特徴を表す情報（例えば紙の色、紙質、用紙の種類等）、（２）文書画像情報として媒体に定着していた物質の特徴を表す情報（例えば筆記具の種類、シミの有無等）、（３）媒体上に表された文書画像情報のイメージとしての特徴を表す情報（例えば余白量、字の種類、筆記者、字の密度、レイアウト等）である。

【０００９】

【作用】本発明によれば、入力された文書画像から文書の外観的な特徴を表す情報を抽出し、これと文書とを対応づけて記憶するため、人間の印象に残り易い文書の外観的な特徴を検索キーとして検索が行える。さらに、このような自然な検索を実現するかなめである文書の外観的な特徴を表す情報は、入力された文書画像から自動的に抽出されるため、特別なセンサは不要であるし、ユーザに余計な負担をかけることもない。

【００１０】

【実施例】以下に、本発明の一実施例を図面を参照して説明する。第１図は、本実施例装置の概略構成図である。ファイルとして格納されるべき文書の書類は、画像入力部１（例えばスキャナ）から画像データとして入力される。次に、特定部分抽出部２が、この入力された画像データに基づいて、書類の構成要素（例えば書類のバックグラウンド部、文字部、表の罫線部、写真部、イラスト部、グラフ部等の種類がある）を抽出する。

【００１１】一方、格納情報生成部３は、入力された画像データをそのままファイルに格納する場合には、画像データをそのままデータ格納部４に格納し、入力された画像データに何らかの処理を施す場合には、画像データをファイルに格納されるデータ・フォーマットに変換してデータ格納部４に格納する。画像データに対し文書構造解析や文字認識処理を施してデータ格納部４に格納しても良い。このように格納されたデータを、格納データと呼ぶ。

【００１２】特定部分特徴識別部５は、特定部分抽出部２で抽出された構成要素について、その特徴（例えば形態、位置、大きさ、形状、色、種類等）を識別する。この特徴の識別は、入力された画像全体に対して、特定部分抽出と並行して行っても良い。

【００１３】派生情報抽出部６は、これら特定部分抽出あるいは特定部分特徴識別の分析結果から、派生情報として用紙の種類、紙質、シミ、紙の色、筆記用具の種類、書き込み比率、書類の種類等を決定する。

【００１４】そして、特定部分抽出部２で得られた構成要素と特定部分特徴識別部５で得られた特徴（構成要素の属性情報と呼ぶ）とが、その他の付属情報と共に格納データに付加されて、それら全体がデータ格納部４にファイルとして格納される。あるいは、派生情報抽出部６で得られた派生情報とその属性情報とが、その他の付属情報と共に格納データに付加されて、それら全体がデータ格納部４にファイルとして格納される。このように格納データに付加される構成要素の情報や派生情報は、文書の外観的な特徴を表す情報である。

【００１５】ファイル検索時には、オペレータにより文書の外観的な特徴を表す情報が検索データとして、検索データ入力部７を介して入力される。検索部８は、格納データに付加された構成要素の情報や派生情報と、入力された検索データとの比較照合を行い、これらが合致した格納データを検索結果として出力部９へ出力する。

【００１６】なお、格納情報生成部３で画像データに対して認識処理を施す場合、認識精度を上げるために前処理としてノイズ除去やシミ抜きを行うことがあるが、ここで除去されるノイズやシミは、特定部分特徴識別部５で識別されるノイズやシミと同じであるから、この前処理部分を両者で共有するようにしても良い。

【００１７】また、入力画像は、カラー（色）データ、グレー（多値）データ、２値データのいずれでも良い。各データに見合った構成要素に関する情報が選択されて、その特徴が識別され格納されることになる。

【００１８】以下に、特定部分抽出部２、特定部分特徴識別部５、派生情報抽出部６について、詳述する。まず、バックグラウンド部を抽出する場合を図２を用いて説明する。入力画像データは色情報により表現されているものとする。

【００１９】まず、入力画像データに対して色分離部で

１１で色分離が行われ、各色毎に分離抽出された画像データが各色画像バッファ１２に記憶される。特定の色のみを色画像バッファに記憶するようにしても良い。色の分離は、ＲＧＢの３原色あるいは明度・彩度・色相の３要素を用いて行われるが、色画像バッファに記憶する段階では、代表色毎（赤、青、黄、緑、紫、橙、藍、白、黒の他、書類に使われる色として水色、ピンク、黄緑等を設定しても良い）に分離されていることが望ましい。この色分離は、原理的には、画像データの各ドットに含まれる色の成分を分析し、そのドットの色がどの代表色に属するかを決定し、決定された代表色に対応する色画像バッファにそのドットの情報を記憶することにより行われる。

【００２０】次に、それらの中で支配的な色をバックグラウンドカラーと決定する。即ち、上記各色画像バッファに記憶されたドット数の合計（総面積）をそれぞれ算出部１３で算出し、この総面積が最大となる色をバックグラウンドカラー決定部１４で決定する。ここで決定された色に基づいて、バックグラウンド部抽出部１５が入力画像中のバックグラウンド部をその他の部分から区別して特定し、バックグラウンドカラーの色画像バッファ中の情報を中心に、バックグラウンド部の情報を色画像バッファから抽出する。このとき、その入力画像データから得られる他の種類の構成要素の特定（後述する文字部等の抽出）を前もってあるいは同時に行い、他の種類の構成要素と特定された部分以外について総面積が最大の部分を抽出するようにすれば、より精度良くバックグラウンド部の抽出が行える。

【００２１】また、総面積によりバックグラウンドカラーを決定するのではなく、入力画像データを各色についてのラン表現に符号化し、各色の部分についてのランの長さやその分布からバックグラウンドカラーを決定することもできる。また、各色毎の連結領域を求め、その連結領域の大きさや、連結領域の面積の平均値や、それらの分布からバックグラウンド部を特定する方法もある。

【００２２】ここで決定されたバックグラウンドカラーは、書類の「紙の色」を表す情報であり、この情報は、「紙の色」抽出部２１（派生情報抽出部６）を介して、データ格納部４にその書類の格納データとともに格納される。また、抽出されたバックグラウンド部の大きさから、書類の「余白の量」を表す情報を抽出し、上記と同様に扱うこともできる。

【００２３】また、ここで抽出されたバックグラウンド部について、ノイズ検出部１７が、バックグラウンド中に含まれる非常に小さい別の色の点の数をカウントし、これの単位面積あたりの密度を計算することにより、バックグラウンドカラー内のノイズの程度を得る。これは、紙の品質（「紙質」）を表す情報であり、この情報は、「紙質」抽出部１９（派生情報抽出部６）を介して、データ格納部４にその書類の格納データとともに格

納される。「紙質」は、ノイズの数値として格納しても良いが、所定のしきい値と比較することにより、「普通紙」「再生紙」「わら半紙」等の情報に変換して格納しても良い。「紙質」の定義は、紙の色や濃度やノイズ量等を総合して決めても良い。また、「紙の厚さ」を検知する機構をスキャナに設けておき、これから得られる情報を上記と同様に扱うこともできる。

【００２４】さらに、バックグラウンドが複数の色から構成されていても良い。このとき、例えば全領域を覆う白い色のバックグラウンド部分と、別の色のより小さい部分のバックグラウンド部分とがあった場合に、白以外のバックグラウンド部分を「シミ部分」なる別の構成要素として「シミ」部分抽出部１６により抽出する。白以外の色で、形状（輪郭）が直線的でない部分を「シミ」部分として決定しても良い。このような「シミ」部分が存在するときには、「シミ」の有無を表す情報を、「シミ」情報抽出部２０（派生情報抽出部６）を介して、データ格納部４にその書類の格納データとともに格納する。この際、「シミ」部分の大きさ（総面積）や位置（中心点あるいは代表点）を、大きさ・位置検出部１８により検出し、「シミ」部分の色とともに「シミ」情報に含めて格納するようにしても良い。これら「シミ」や「紙質」の抽出は入力画像がグレーでもできる。

【００２５】次に、文字部を抽出する場合を図３を用いて説明する。入力画像データはカラーでもグレーでも２値でも良い。まず、連結領域抽出部３１で、入力画像データから、ある程度黒画素が固まって存在する連結領域（大抵の場合１文字が１連結領域を構成する）を抽出する。そして解析部３２で、連結領域、または入り組んでいる連結領域、またはごく近くに存在する連結領域をマージしてできる領域について、それらの並びが直線的であるか否か、それらの大きさが揃っているか否か、並びのピッチがほぼ一定であるか否か、あるいはそれらに対して文字認識を行った結果妥当な確信度が得られたか否か等を判断し、文字部抽出部３３で、抽出された連結領域を文字部として特定するか否かを決定する。

【００２６】ここで抽出された文字部の特徴を以下のように識別する。まず、カラー画像が入力された場合には、文字の色検出部３５で文字部の色を決定する。１枚の書類に異なる色の文字部が複数存在する場合、それぞれの文字部について色を検出する。そして検出された文字の色を表す情報を、データ格納部４にその書類の格納データとともに格納する。複数色がある場合には、文字部の位置と色とを対応させた情報を格納する。

【００２７】また、文字部の特徴の１つである文字タイプには、手書、活字、はんこう、ドット印字、フォントの種類等がある。これらの種類は、その文字の大きさ、色、文字列の並び方、文字列を囲む枠の形状等により識別されて、その結果がデータ格納４に格納される。例えば、人間の記憶に残り易い特徴としては、書類の大部分

が手書の文字で書かれていたか、印刷あるいはプリントアウトされた活字であったかが挙げられることに着目し、文字タイプ判定部３７で、手書辞書４０あるいは活字辞書４１の少なくとも一方を用いて、手書か活字かを識別することとする。例えば活字辞書４１を用いて文字認識を行えば、書類の大部分が手書の文字であれば低い確信度しか得られない文字が多く、活字であればほぼ妥当な確信度が得られるため、確信度の全文字についての合計が所定のしきい値より高ければ活字、低ければ手書と判断する。この手書か活字かの情報は文字タイプ情報として、データ格納部４にその書類の格納データとともに格納される。なお、手書か活字かは、辞書を用いずに、連結領域の縦方向、横方向の並びにおいて、そのずれが非常に小さい場合には活字であると判断し、バラバラであれば手書と判断することもできる。

【００２８】ここで手書と判断された場合、さらに次の処理が可能である。即ち、本装置をパーソナルに使用するとすれば、自分が書いたものか、他人が書いたものかが重要となる。そこで、本装置の所有者の手書文字の特徴パターンを有する辞書４５を用いて、その確信度の高低により、筆記者識別部４４が筆記者が所有者であるか否かを判断する。辞書４５を複数人について持てば、その書類の筆記者の名前を推定することもできる。この筆記者情報は、データ格納部４にその書類の格納データとともに格納される。

【００２９】また、手書の場合、筆記用具の種類を特定することもできる。即ち、筆記具識別部４６で、文字線のかすれ方や濃度（画像データはグレーであることが必要）や、文字線の太さ検出部３６により検出される線幅等から、鉛筆で書かれたものか、ボールペンで書かれたものか、サインペンで書かれたものか等を判定する。この筆記具情報は、データ格納部４にその書類の格納データとともに格納される。文字線の太さをそのまま筆記具情報として格納しても良い。なお、画像入力部１に反射光の検出器を別に用意し、反射率や分光特性を解析する手段を付加し、紙の上に定着している物質が、鉛筆の芯か、ボールペンのインクか、サインペンのインクか、コピーのトナーか、あるいはプリンタのリボンか等を識別することにより、筆記用具の種類あるいはコピーと原紙の区別を特定するようにすることもできる。

【００３０】活字と判断された場合には、そのフォント（明朝体、毛筆、ゴシック、イタリック等）をフォント識別部４７でさらに判別し、データ格納部４に格納しても良い。これらの特徴を表す情報は、書類の大部分を占める文字についてのものが抽出されれば足りる。

【００３１】また、手書文字と活字文字とが混在していると判断される場合には、混在率検出部４８により、手書文字の推定文字数、あるいはその推定手書文字数の推定全文字数に対する比を算出し、書類中に手書で書き込んだ文字の量を表す情報として、データ格納部４に格納

することも有効である。この情報は、手書文字の存在する領域と活字文字の存在する領域との総面積の比を算出して求めることもできる。

【００３２】文字部の特徴の他の１つである文字種には、数字、英字、カナ、漢字等がある。ここで、人間の記憶に残り易い特徴としては、書類が英語で書かれていたか、日本語で書かれていたかが挙げられることに着目し、文字種判定部３８で、英字辞書４２あるいはカナ漢字辞書４３の少なくとも一方を用いて、英語か日本語かを識別することとする。例えば英字（アルファベット）辞書４２を用いて文字認識を行えば、書類が日本語であれば低い確信度しか得られない文字が多く、英語であればほぼ妥当な確信度が得られるため、確信度の全文字についての合計が所定のしきい値より高ければ英語、低ければ日本語と判断する。この言語種類の情報は、データ格納部４にその書類の格納データとともに格納される。さらに、数字についても同様な処理を行い、帳票のように数字が羅列された書類であることを示す情報を上記言語種類の情報に加えることもできる。

【００３３】他に、抽出された文字部に対して、ピッチ検出部３９が文字ピッチや行ピッチを検出し、この結果を元に縦横識別部４９が、その書類が縦書きであるか横書きであるかを識別する。これには４つの状態が有り得、１つ目は例えばＡ４の用紙を縦に置いて横書きしたもの、２つ目は用紙を横に置いて縦書きしたもの、３つ目は用紙を縦に置いて縦書きしたもの、４つ目は用紙を横に置いて横書きしたものである。そこで、横方向のピッチが縦方向のピッチより小さければ上記１か２であると判定し、逆ならば３か４と判定する。さらに、用紙を置いた向きそのままで読めるように文字が書かれていると仮定した文字認識と、用紙の向きを直角方向に置き換えた場合に読めるように文字が書かれていると仮定した文字認識の双方を行って、結果を比較することにより、１と２の区別、あるいは３と４の区別を行う。この４つの状態のいずれであるかを示す情報は、データ格納部４にその書類の格納データとともに格納される。

【００３４】また、文字の大きさ・密度検出部５０が文字の大きさや密度を判定することもできる。この場合も、大きさや密度を示す数値をそのままデータ格納部４に格納するのではなく、「細かい字・びっしり」「大きい字・すかすか」のような情報に変換して格納しても良い。

【００３５】上記の例ではバックグラウンド部から「紙質」や「シミ」部分を抽出する処理を説明したが、文字部に対して文字認識を施す際に、通常認識率を上げるために前処理として行う正規化等を行わないで、そのままのデータに対して認識を行うことにより、認識率がまとまって所定のしきい値より悪い部分を「シミ」部分と特定したり、認識率が１枚の紙全体に対して悪いならば「質の悪い紙」と特定したりすることもできる。

【００３６】以下にその他の特定部分（構成要素）を抽出する場合を図４を用いて説明する。表の罫線部の抽出は、直線・曲線検出部６１により直線が数多く検出され、交わり検出部６２により前記の直線が互いに直行している交差点が数多く検出され、解析部６３により各直線の位置や長さが揃っていると判断されるエリアを、表の罫線部として抽出することにより行われる。

【００３７】その後、用紙の種類判定部７１で、一定ピッチで並ぶ直線が紙面全体に存在すると判断される場合には、用紙の種類が罫線入りレポート用紙あるいは便箋であると決定する。さらに直線の並びにより、縦罫線の用紙か横罫線の用紙かをも決定できる。また、よく使う用紙の罫線の並びや色や印（社名入り等）を用紙種辞書７２に登録しておき、抽出された直線群等とのマッチングをとることにより、用紙の種類を「自社製レポート用紙」「Ａ部課提出用記入用紙」のように特定することができる。この用紙の種類を示す情報や、表の罫線部の位置・大きさを表す情報は、データ格納部４にその書類の格納データとともに格納される。

【００３８】図面部の抽出は、直線や曲線が数多く検出され、それらの交差点が数多く検出され、それらが表の罫線部とみなされないエリアを抽出することにより行われる。

【００３９】写真部は、画像処理技術として知られている像域分離技術を用いて抽出できる。写真部には、画像の濃淡が滑らかに変化するグラビア写真部と、画像の部分に応じてその大きさが変動する黒点が並んでいることが特徴である網掛写真部とがある。また、写真部の色を分析することにより、カラー写真かモノクロ写真かの判定ができる。

【００４０】グラフ部の抽出は、図面認識で通常使われている円抽出や矩形抽出、線分抽出等の技術を使うことにより実現される。これらの抽出処理は、前述のように抽出された図面部にのみ行うことにより、図面がグラフであるかその他の図面であるかを特定するようにしても良い。グラフ部には、棒グラフと円グラフと折れ線グラフとがある。

【００４１】このように抽出されたバックグラウンド、文字、図面、写真、グラフ等の構成要素は、その位置や大きさ、さらに種類等の属性情報、派生情報も含めて、データ格納部４にその書類の格納データとともに格納される。このとき、位置や大きさそのものを格納するのではなく、各構成要素の位置関係・比率検出部７３を介して、「右上に写真部が、左下にグラフ部が存在する」のような位置関係の情報や、「図面が全体の６割を占めている」、「図面と文字が１：２の比率で存在する」のような比率の情報に変換して、格納することも有効である。

【００４２】別の構成要素として、予め指定した場所に存在する印や色も考えられる。即ち指定された場所に特定の印あるいは色が存在するか否かを検出して、この情報をデータ格納部４にその書類の格納データとともに格納する。例えばユーザが重要と思う書類にはその右上隅に赤ペンでチェックをしておくことにすると、入力された画像の右上隅に赤い色が存在するか否かを検出して、重要な書類か否かを表す情報として格納データに付加すると効果的である。また場所を特定せず、全画面をサーチしてその特定の印あるいは色を発見するようにしても良い。

【００４３】以上説明した情報がどのように格納されるかを図５に示す。各構成要素や派生情報、それらの属性情報には、名称に対応する数値データあるいはコードを割り当てる。構成要素については、図の左半分に示すように、属性名とこれの値である属性値とを組にし、これを属性セットと呼ぶ。この属性セット（複数）と構成要素とを組にし、例えば表形式で、格納する。派生情報については、図の右半分に示すように、派生情報とその属性値とを組にして格納する。これらの一方だけを格納しても本発明の効果は得られる。また図５は例示であり、これら全ての情報を格納する必要はない。

【００４４】上記の構成要素についての表形式の情報、派生情報とその属性値との組で表される情報は、格納データ（文書画像やその認識結果）とは別の場所、例えばディレクトリ部に格納しても良いし、格納データのヘッダ部分に付加して格納しても良い。別に格納した方が、これらの情報を用いて検索する場合に、ディレクトリ部のみを検索し、合致したものについてのみ格納データを読み出せば良いので、検索速度は早くなる。

【００４５】また、各構成要素の中に含まれる属性名の種類、派生情報の種類、派生情報で定義できる属性値の種類は、予め定めておく。つまり、例えば図面部であれば、これについての属性名は、色と大きさと位置の３種類のようにである。そして、予め表のどこ（メモリのアドレス）にどの構成要素のどの属性名を割り当てるか決めておく。そして、特定部分抽出部２や特定部分特徴識別部５で求められた属性値を、対応する属性名のところに書き込む。抽出や識別に失敗したり、スキャナがカラーでなく色は求められないような場合には、求めることができなかった属性名のところにＮＵＬＬを書き込む。

【００４６】派生情報についても、予めメモリのどの格納位置にどの派生情報を割り当てるか決めておく。そして、各派生情報について定義できる属性値も、例えば余白であれば多・中・少の３種類、文字タイプであれば手書・活字の２種類のように、予め決められている。この各派生情報について定義されている属性値は、テーブルの形で記憶しておくと、後で述べる検索の際に便利なことがある。派生情報抽出部６で求められる属性値は、予め定義されている中から選ばれるものであり、この求められた属性値を、対応する派生情報のところに書き込む。求めることができなかった派生情報のところにはＮ

ＵＬＬを書き込む。

【００４７】検索時の動作の一例を図６（ａ）に示す。検索データ入力部７からは、例えば「ピンクの紙に自分で書いたもので、コーヒーのシミがついている文書」のように自然言語で入力する。すると、検索情報抽出部８１は、各派生情報とこれについて予め定義されている属性値を対応させた表を記憶している記憶部８２の情報を用いて、上記の検索データから「ピンク」「自分」「シミ」という検索情報の元となるワードを抽出し、「紙の色：ピンク」「筆記者：自分」「シミ：有」という３つの検索情報を得る。そして、検索情報比較照合部８３が、得られた検索情報の項目（「紙の色」等）を含む属性セット（図５右半分のような派生情報と属性値の組）をデータ格納部４に各格納データ（文書）に付随して格納されている中から探し、この属性値と検索情報のそれ（「ピンク」等）とを比較照合し、これらが合致する文書を選択して、文書提示部８５へ出力する。

【００４８】合致するかどうかの判断においては、属性値同士の類似度を定義しておき（例えば完全一致は類似度１００％、ピンクと赤は類似度が８０％、白と黒は類似度０％、格納されている属性値がＮＵＬＬであれば類似度は５０％（判断できないことを示す）等）、検索情報抽出部８１で得られた検索情報の全てが完全一致でなくとも、各検索情報についての比較照合の結果である類似度を全検索情報で合計した値が所定のしきい値より大きければその文書を選択するようにしても良い。また、派生情報の中には、「紙の色」のようにある程度の確信度を持って属性値を抽出できる性質のものと、「筆記者」のように自分なのか他人なのかの決定に曖昧さが残ることが多い性質のものとがある。そこで、派生情報毎に重みを予め定めておき、前記の類似度合計の際に、確からしい派生情報についての類似度を重視し、曖昧な派生情報についての類似度は参考程度にするように、重み付けした合計を行うようにしても良い。この重みは、予め定めておくのではなく、派生情報抽出部６での抽出の際に確信度をも求めることにより決定しても良い。

【００４９】また、通常のファイリング装置におけるキーワード検索を併用するのも有効である。つまり、文書の内容を表すキーワードをファイル時に文書データに自動あるいは手動で付加しておき、検索時にこのキーワードが思い出せればまずキーワードで検索件数を絞り込み、その後上記のような派生情報を用いた検索を行う。このようにすれば、キーワードを付加する際に、そのキーワードがユニークなものであるか否かについて注意する必要がなくなり、ユーザの負担を軽減できる。その他、「何月頃ファイルした」という情報を、ファイリング装置に備えた時計機能で取り出して文書データに付加しておき、この時間に関する情報と上記のような派生情報とを組み合わせて用いて検索しても、検索精度を上げることができる。

【００５０】上記では派生情報を用いた検索を説明したが、構成要素とその属性情報を用いた検索は以下のようにできる。この場合、属性値は派生情報の場合よりも生データに近いものが格納されているので、記憶部８２には各構成要素とそれが持つ属性名、及びその組が表す情報名を記憶しておく。そして、例えば「紙の色はピンクで、シミがついていて、シミの大きさは大きく、シミの位置は右上あたりだった文書」という検索データが入力されたとすると、記憶部８２の情報名とマッチングを取りながら「紙の色」「シミの大きさ」「シミの位置」という検索情報の項目を抽出し、各項目の直後に書かれた検索データである「ピンク」「大」「右上」を抽出し、それぞれセットとして検索情報とする。

【００５１】さらに、記憶部８２の情報を用いて、「紙の色」という情報名を「バックグラウンド部のカラー」という構成要素と属性名に変換した後、検索情報と、各文書に付随して格納されたデータ格納部４の図５左半分のような構成要素と属性情報の組との比較照合を行う。このとき、まず「バックグラウンドの部のカラー」という検索情報の項目を含む属性セットをデータ格納部４の中から探し、この属性値と検索情報のそれ（「ピンク」等）とを比較照合し、これらが合致する文書を選択する。検索情報では「大」のように大まかな表現がされているが、例えば「シミの大きさ」については数値「１～１０」が「小」、「１１～２０」が「中」、「２１～３０」が「大」のような対応を予め記憶しておくことにより、「大」であれば「２１～３０」という数値に変換して、データ格納部４の属性値との比較照合を行う。この場合は、数値同士の比較照合であるから、類似度計算は簡単にできる。

【００５２】尚、派生情報と構成要素の属性情報の双方を用いた検索もできる。特に、派生情報に「シミ：有」のような大まかな情報が、構成要素の属性情報に「シミ」の色、大きさ等の細かい情報が入っている場合、「シミのついた」という検索データからまず文書データに付加された派生情報を見て「シミ：有」の文書（複数）を選択した後、対応する構成要素「シミ」の属性名である「色」「大きさ」等を提示し、ユーザが「色」は「茶」、「大きさ」は「大」のように検索データの続きを入力して、構成要素の属性情報による絞り込みを行う。また、派生情報の中にも、「文字タイプ」と「フォント」あるいは「筆記者」のように、「文字タイプ」が「活字」なら「フォント」という派生情報があり得る一方、「手書」なら「筆記者」という派生情報があり得るというように階層構造を持つものがあり、ここでも前述した対話的な絞り込みが可能である。

【００５３】図６（ｂ）には、検索時の動作の別の例を示す。まず、派生情報項目表示部８６が、派生情報抽出部６により抽出可能な派生情報を、それについて予め定義されている属性値（複数）とともに、図中１００のよ

うに表示する。ユーザはこれを見て、検索データ入力部７により、所望の文書の「紙の色」は「ピンク」だった、「余白」は中くらいだった、のように思い出しながら、各派生情報について指示していく。思い出せない場合には、その項目を除いて後の比較照合を行うので、その項目については入力しなくて良い。このように入力された検索データに対して、検索情報比較照合部８３が、図６（ａ）の場合と同様に合致する文書を選択、提示する。

【００５４】以上は、ファイルされる文書画像から構成要素の属性情報や派生情報を抽出して、これを用いて検索を行う実施例であるが、これらをパラパラめくりに用いることも有効である。つまり、格納データ（文書）をパラパラめくりながら提示することによりユーザに所望の文書を選択させるシステムにおいて、提示する文書に付随して格納されている構成要素の属性情報や派生情報を画像に展開する。例えば、格納されている「シミ」の色や大きさの情報に従ってその文書の画像に「シミ」の画像情報を重畳して表示する。これにより、特に、入力された文書画像からノイズを除去したものが格納データとした格納される場合には、パラパラめくりのとき提示される文書に見覚えのあるノイズがないためにユーザが一見して所望の文書か否かを判断することができないということがなくなり、使い勝手を向上させることができる。

【００５５】また、原文書画像あるいは格納データと共に、抽出された派生情報や属性情報（例えば紙の色：ピンクのように表示）し、派生情報や属性情報をユーザが修正変更、追加できるよう構成しても良い。この場合、ユーザが例えば紙の色：白と修正し、書込量：多という情報を追加し、シミ：有りという情報を削除したとすると、このように修正された派生情報等を該当文書に対応付けてデータ格納部４に格納する。

【００５６】

【発明の効果】以上詳述したように、本発明によれば入力画像から自動的に抽出される文書の外観を表す情報（紙の色、紙質、シミ、文字の色、文字タイプ、文字種、筆記者、筆記具、フォント、余白への書き込みの量、書類の縦横、文字の大きさ・密度、用紙の種類、図面や写真の位置関係等）を用いて、所望の文書が検索でき、ユーザが文書の内容やキーワードを明確に覚えていない場合にも、その文書の周辺的な情報を思い出すことによる検索が実現できる。

【図面の簡単な説明】

【図１】　本実施例装置の概略構成を示す図。

【図２】　本実施例装置でバックグラウンド部を抽出する場合の処理例を示す図。

【図３】　本実施例装置で文字部を抽出する場合の処理例を示す図。

【図４】　本実施例装置で表の罫線部、図面部、写真部、グラフ部を抽出する場合の処理例を示す図。

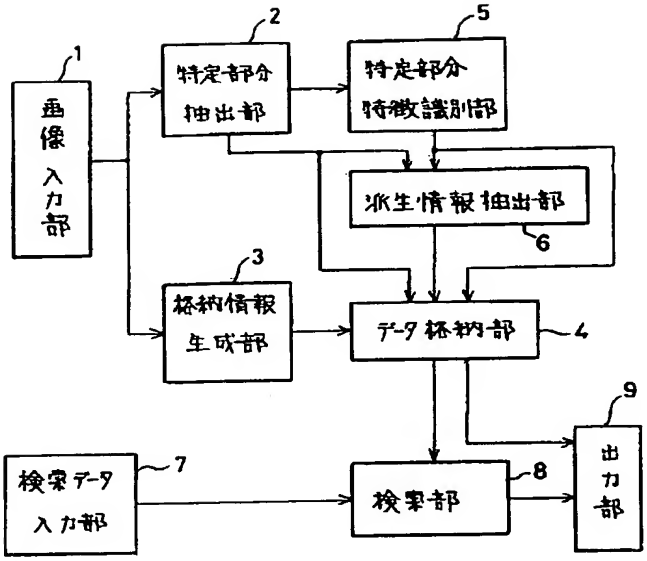
【図５】　データ格納部４に格納される構成要素の情報や派生情報の形式例を示す図。

【図６】　本実施例装置における検索のための構成を示す図。

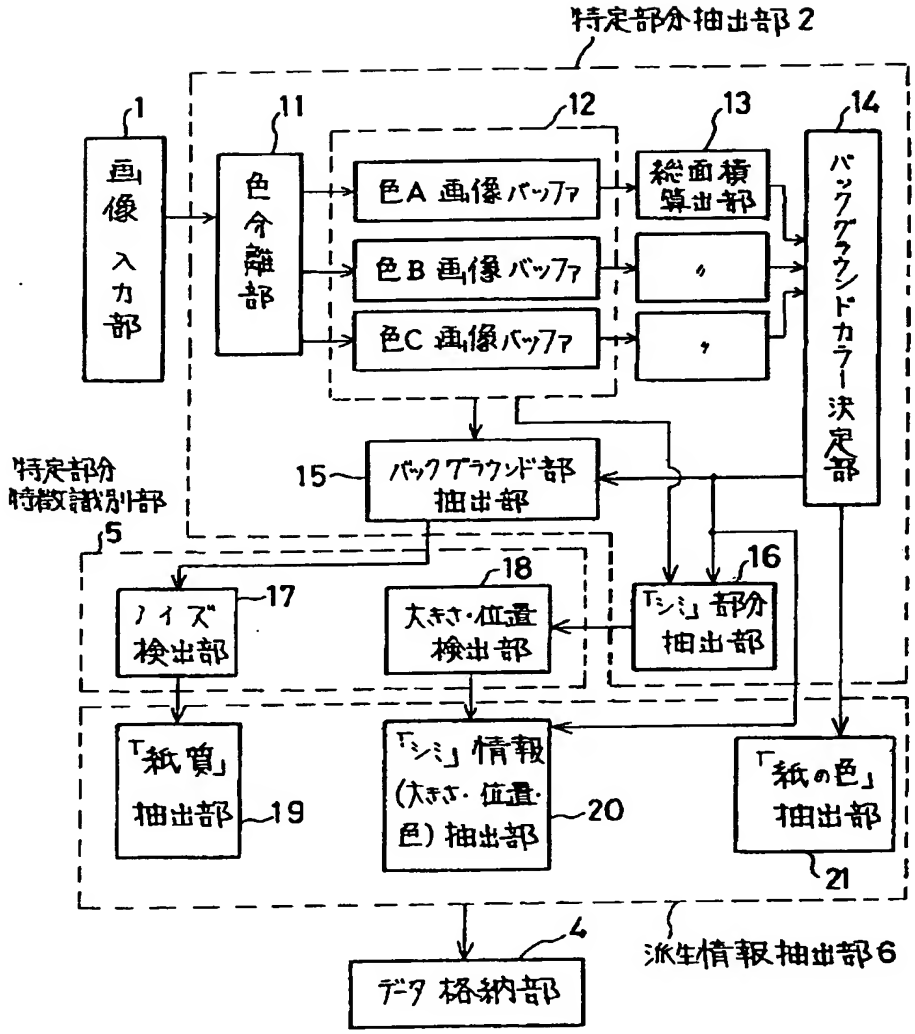
【符号の説明】

１…画像入力部、２…特定部分抽出部、３…格納情報生成部、４…データ格納部、５…特定部分特徴識別部、６…派生情報抽出部、７…検索データ入力部、８…検索部、９…出力部、１１…色分離部、１２…色画像バッファ、１３…総面積算出部、１４…バックグラウンドカラー決定部、１５…バックグラウンド部抽出部、１６…「シミ」部分抽出部、１７…ノイズ検出部、１８…大きさ・位置検出部、１９…「紙質」抽出部、２０…「シミ」情報抽出部、２１…「紙の色」抽出部、３１…連結領域抽出部、３２…解析部、３３…文字部抽出部、３４…画像バッファ、３５…文字の色検出部、３６…文字線の太さ検出部、３７…文字タイプ判定部、３８…文字種判定部、３９…ピッチ検出部、４０…手書辞書、４１…活字辞書、４２…英字辞書、４３…カナ漢字辞書、４４…筆記者識別部、４５…所有者手書辞書、４６…筆記具識別部、４７…フォント識別部、４８…混在率検出部、４９…縦横識別部、５０…文字の大きさ・密度検出部、６１…直線・曲線検出部、６２…交わり検出部、６３・６５…解析部、６４…表の罫線部抽出部、６６…図面部抽出部、６７…像域分離部、６８…写真部抽出部、６９…円・矩形・線分抽出部、７０…グラフ部抽出部、７１…用紙の種類判定部、７２…用紙種辞書、７３…各構成要素の位置関係・比率検出部、８１…検索情報抽出部、８２…派生情報・属性値対応表記憶部、８３…検索情報比較照合部、８４…類似度・重み記憶部、８５…文書提示部、８６…派生情報項目表示部

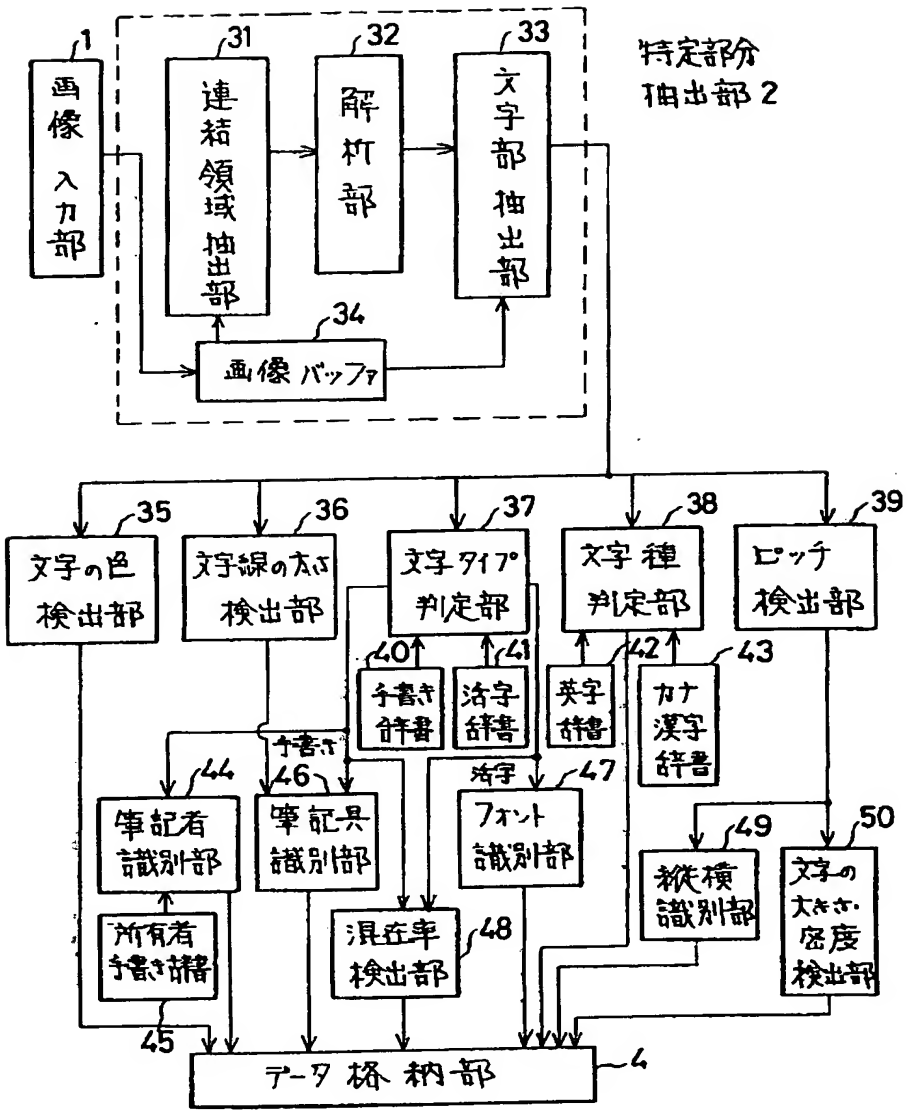
【図１】


1
画像入力部
2
特定部分抽出部
5
特定部分特徴識別部
派生情報抽出部
6
3
格納情報生成部
データ格納部
4
9
出力部
検索データ入力部
7
検索部
8

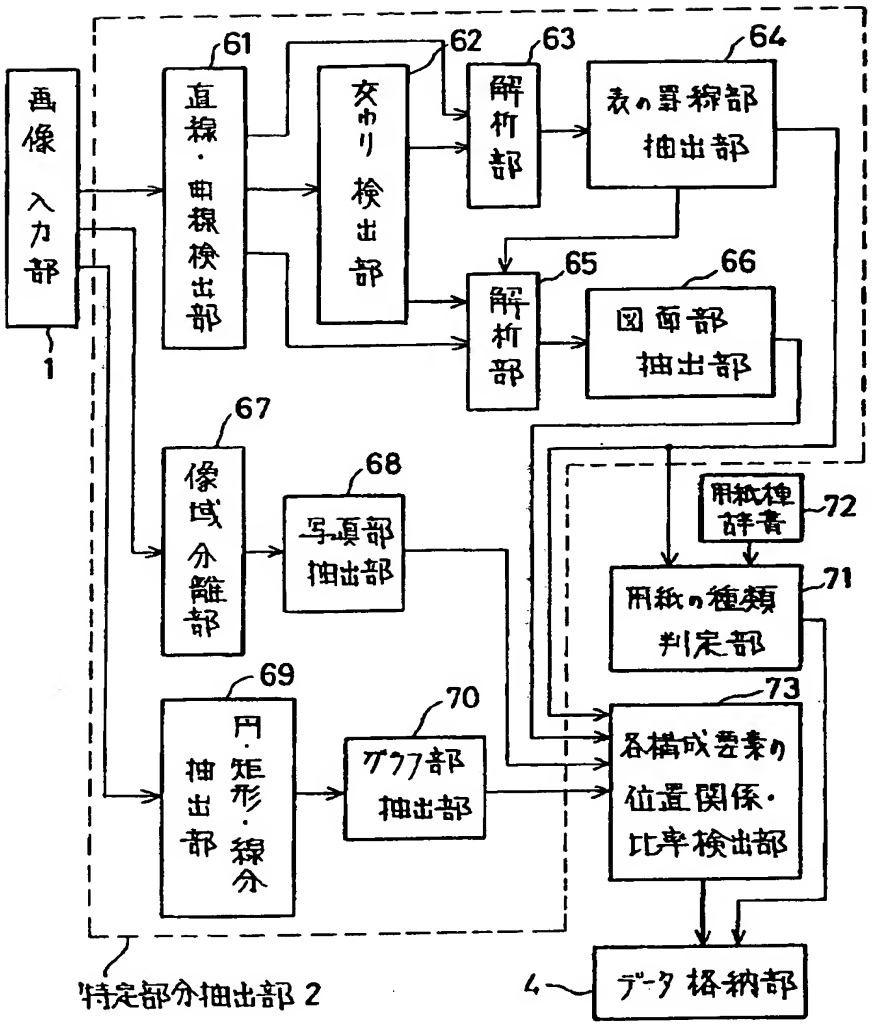
【図２】


特定部分抽出部２
1
画像入力部
11
色分離部
12
色A画像バッファ
色B画像バッファ
色C画像バッファ
13
総画素数算出部
〃
〃
14
バックグラウンドカラー決定部
特定部分特徴識別部
5
15
バックグラウンド部抽出部
16
「シミ」部分抽出部
17
ノイズ検出部
18
大きさ・位置検出部
「紙質」抽出部
19
「シミ」情報（大きさ・位置・色）抽出部
20
「紙の色」抽出部
21
派生情報抽出部６
4
データ格納部

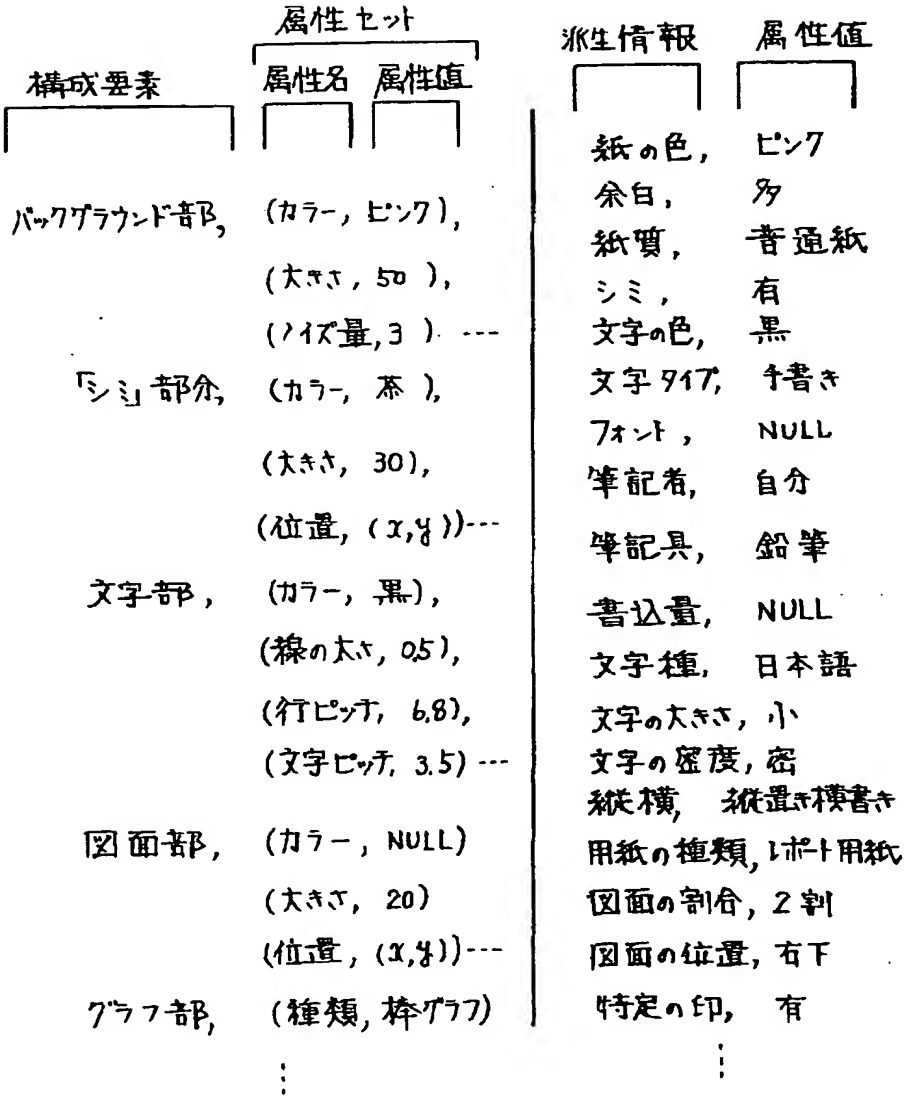
【図３】


1
画像入力部
31
連結領域抽出部
32
解析部
33
文字部抽出部
特定部分抽出部２
34
画像バッファ
35
文字の色検出部
36
文字線の太さ検出部
37
文字タイプ判定部
38
文字種判定部
39
ピッチ検出部
40
手書き辞書
41
活字辞書
42
英字辞書
43
カナ漢字辞書
44
筆記者識別部
所有者手書き辞書
45
46
筆記具識別部
47
フォント識別部
48
混在率検出部
49
縦横識別部
50
文字の大きさ・密度検出部
データ格納部
4

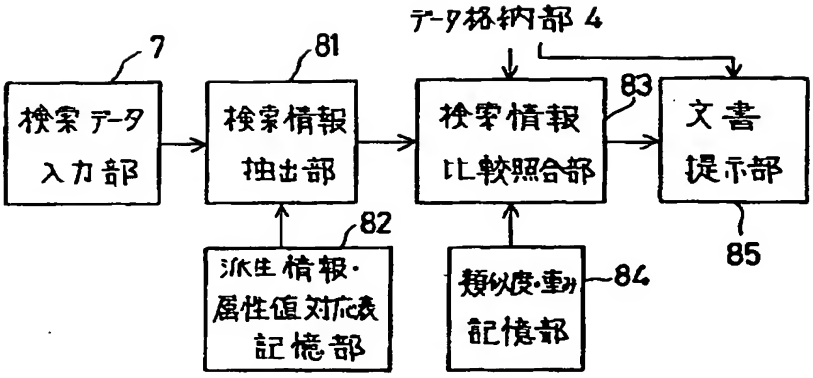
【図4】


画像入力部
1
直線・曲線検出部
61
交わり検出部
62
解析部
63
表の罫線部抽出部
64
解析部
65
図面部抽出部
66
像域分離部
67
写真部抽出部
68
円・矩形・線分抽出部
69
グラフ部抽出部
70
用紙種辞書
72
用紙の種類判定部
71
各構成要素の位置関係・比率検出部
73
特定部分抽出部 2
データ格納部
4

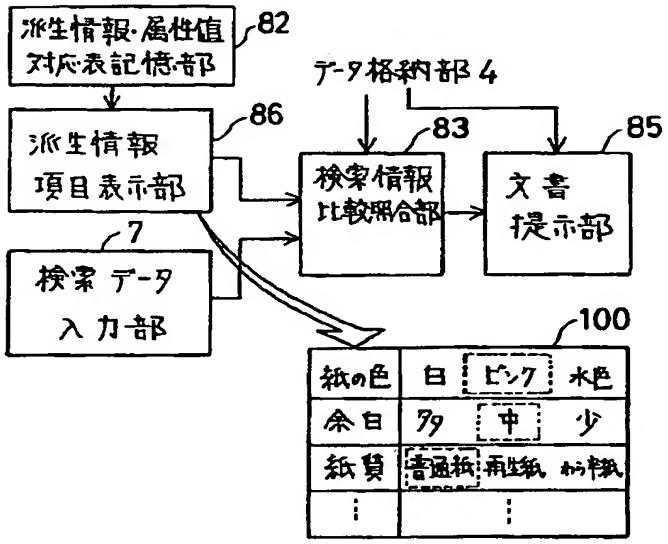
【図5】


属性セット
構成要素
属性名
属性値
派生情報
属性値
バックグラウンド部,
(カラー, ピンク),
(大きさ, 50),
(ノイズ量, 3) ---
「シミ」部分,
(カラー, 茶),
(大きさ, 30),
(位置, (x,y)) ---
文字部,
(カラー, 黒),
(線の太さ, 0.5),
(行ピッチ, 6.8),
(文字ピッチ, 3.5) ---
図面部,
(カラー, NULL)
(大きさ, 20)
(位置, (x,y)) ---
グラフ部,
(種類, 棒グラフ)
紙の色, ピンク
余白, 多
紙質, 普通紙
シミ, 有
文字の色, 黒
文字タイプ, 手書き
フォント, NULL
筆記者, 自分
筆記具, 鉛筆
書込量, NULL
文字種, 日本語
文字の大きさ, 小
文字の密度, 密
縦横, 縦置き横書き
用紙の種類, レポート用紙
図面の割合, 2割
図面の位置, 右下
特定の印, 有

【図６】


データ格納部４
7
81
83
検索データ入力部
検索情報抽出部
検索情報比較照合部
文書提示部
85
82
派生情報・属性値対応表記憶部
類似度・重み記憶部
84

（ａ）


派生情報・属性値対応表記憶部
82
データ格納部４
86
83
85
派生情報項目表示部
検索情報比較照合部
文書提示部
7
検索データ入力部
100
紙の色
白
ピンク
水色
余白
多
中
少
紙質
普通紙
再生紙

（ｂ）